TILBURG UNIVERSITY

Tilburg University

Multivariate Student -t Regression Models

Fernández, C.; Steel, M.F.J.

Publication date:
1997

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Fernández, C., & Steel, M. F. J. (1997). *Multivariate Student -t Regression Models: Pitfalls and Inference*. (CentER Discussion Paper; Vol. 1997-08). Econometrics.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

Multivariate Student-t Regression Models: Pitfalls and Inference

By Carmen Fernández and Mark F.J. Steel [1]

CentER for Economic Research and Department of Econometrics
Tilburg University, 5000 LE Tilburg, The Netherlands

FIRST VERSION DECEMBER 1996; CURRENT VERSION JANUARY 1997

Abstract

We consider likelihood-based inference from multivariate regression models with independent Student-t errors. Some very intruiging pitfalls of both Bayesian and classical methods on the basis of point observations are uncovered. Bayesian inference may be precluded as a consequence of the coarse nature of the data. Global maximization of the likelihood function is a vacuous exercise since the likelihood function is unbounded as we tend to the boundary of the parameter space. A Bayesian analysis on the basis of set observations is proposed and illustrated by several examples.

KEY WORDS: Bayesian inference; Coarse data; Continuous distribution; Maximum likelihood; Missing data; Scale mixture of Normals.

1. INTRODUCTION

The multivariate regression model with unknown scatter matrix is widely used in many fields of science. Applications to real data often indicate that the analytically convenient assumption of Normality is not quite tenable and thicker tails are called for in order to adequately capture the main features of the data. Thus, we consider regression error vectors that are distributed as scale mixtures of Normals. We shall mainly emphasize the empirically relevant case of independent sampling from a multivariate Student-t distribution with unknown degrees of freedom. In particular, we provide a complete Bayesian

[1] Carmen Fernández is Research Fellow, CentER for Economic Research and Assistant Professor, Department of Econometrics, Tilburg University, 5000 LE Tilburg, The Netherlands. Mark Steel is Senior Research Fellow, CentER for Economic Research and Associate Professor, Department of Econometrics, Tilburg University, 5000 LE Tilburg, The Netherlands. We gratefully acknowledge the extremely valuable help of F. Chamizo in the proof of Theorem 3 as well as useful comments from B. Melenberg and W.J. Studden. Both authors benefitted from a travel grant awarded by the Netherlands Organization for Scientific Research (NWO) and were visiting the Statistics Department at Purdue University during much of the work on this paper.

analysis of the linear Student-t regression model, and also comment on the behaviour of the likelihood function.

The Bayesian model will be completed with a commonly used improper prior on the regression coefficients and scatter matrix, and some proper prior on the degrees of freedom. Section 3 examines the usual posterior inference on the basis of a recorded sample of point observations. Even though Theorem 1 indicates that Bayesian inference is possible for almost all samples (*i.e.* except for a set of zero probability under the sampling model), problems can occur since any sample of point observations formally has probability zero of being observed. In practice, this can become relevant due to rounding or finite precision of the recorded observations, and we can easily end up with a sample for which inference is precluded. This incompatibility between the continuous sampling model and any sample of point observations can have very disturbing consequences: the posterior distribution may not exist, even if it already existed on the basis of a subset of the sample. New observations can, thus, have a devastating effect on the usual Bayesian inference. Fernández and Steel (1996a) present a detailed discussion of this phenomenon in the context of a univariate location-scale model.

Section 4 presents a solution through the use of set observations, which have positive probability under the sampling model, and are, thus, in agreement with the sampling assumptions. This leads to a fully coherent Bayesian analysis where new observations can never harm the possibility of conducting inference. A Gibbs sampling scheme [see *e.g.* Gelfand and Smith (1990) and Casella and George (1992)] is seen to be a convenient way to implement this solution in practice. Some examples are presented: a univariate regression model for the well-known stackloss data [see Brownlee (1965)], and a bivariate location-scale model for the *iris setosa* data of Fisher (1936).

The analysis through set observations is naturally extended to the case where some components of the multivariate response are not observed (missing data). We illustrate this with the artificial Murray (1977) data, extended with some extreme values in Liu and Rubin (1995).

In addition, we find that none of the results concerning the feasibility of Bayesian inference with set observations depend on the particular scale mixture of Normals that we sample from.

Finally, in Section 5 the Student likelihood function for point observations is analyzed in some detail: it is found that the likelihood is unbounded as we tend to the boundary of the parameter space in a certain direction. This casts some doubt on the meaning and validity of a maximum likelihood analysis of this model [as performed in *e.g.* Lange, Little and Taylor (1989), Lange and Sinsheimer (1993) and Liu and Rubin (1994, 1995)]. This behaviour of the likelihood function is illustrated through the stackloss data example, and it also explains the source of the problems encountered by Lange *et al.* (1989) and Lange and Sinsheimer (1993) when applying the EM algorithm for joint estimation of regression coefficients, scale and degrees of freedom to the radioimmunoassay data set of Tiede and Pagano (1979).

All proofs are grouped in Appendix A, whereas Appendix B recalls some matricvariate probability densities used in the body of the paper. With some abuse of notation, we do not explicitly distinguish between random variables and their realizations, and $p(\cdot)$ (a density

function) or $P(\cdot)$ (a measure) can correspond to either a probability measure or a general σ-finite measure. All density functions are Radon-Nikodym derivatives with respect to the Lebesgue measure in the corresponding space, unless stated otherwise.

2. THE MODEL

Observations for the p-variate response variable y_i are assumed to be generated through the linear regression model

$$y_i = \beta' x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where β is a $k \times p$ matrix of regression coefficients, x_i is a k-dimensional vector of explanatory variables and the entire design matrix, $X = (x_1, \ldots, x_n)'$, is taken to be of full column rank k [denoted as $r(X) = k$]. The error vectors ε_i are independent and identically distributed (i.i.d.) as p-variate scale mixtures of Normals with mean zero and positive definite symmetric (PDS) covariance matrix Σ. The mixing variables, denoted by λ_i, $i = 1, \ldots, n$, follow a probability distribution $P_{\lambda_i|\nu}$ on $\Re_+$, which can depend on a parameter $\nu \in \mathcal{N}$ (possibly of infinite dimension). Thus, we have n independent replications from the sampling density

$$p(y_i|\beta, \Sigma, \nu) = \int_0^\infty \frac{\lambda_i^{p/2}}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{\lambda_i}{2}(y_i - \beta' x_i)'\Sigma^{-1}(y_i - \beta' x_i)\right\} dP_{\lambda_i|\nu}. \tag{2.2}$$

By changing $P_{\lambda_i|\nu}$ we cover the class of p-variate scale mixtures of Normals. The latter is a subset of the elliptical class [see Fang, Kotz and Ng (1990), chap. 2], with ellipsoids in $\Re^p$ as isodensity sets, while allowing for a wide variety of tail behaviour. Leading examples are finite mixtures of Normals, corresponding to a discrete distribution on λ_i, and multivariate Student-t distributions with $\nu > 0$ degrees of freedom, where $P_{\lambda_i|\nu}$ is a Gamma distribution with unitary mean and both shape and precision parameters equal to $\nu/2$. Most of the subsequent discussion will focus on the empirically relevant case of Student-t sampling.

Special cases of the model in (2.1) are the multivariate location-scale model, where $k = 1$ and $x_i = 1$, and the univariate regression model for $p = 1$.

The Bayesian model needs to be completed with a prior distribution for (β, Σ, ν). In particular, we assume a product structure between the three parameters where

$$p(\beta, \Sigma) \propto |\Sigma|^{-(p+1)/2} \tag{2.3}$$

and

$$P_\nu \text{ is any probability measure on } \mathcal{N}. \tag{2.4}$$

The prior in (2.3) is the "usual" default prior in the absence of compelling prior information on (β, Σ). Under fixed ν it corresponds to the Jeffreys' prior under "independence" and is thus invariant under separate reparameterizations of β and of Σ.

Note that the model in (2.1) implies that all p components of y_i are regressed on the same variables x_i. Thus, we treat a special case of Zellner's (1962) seemingly unrelated

regression (SUR) model, which allows for different regressors on the p components. Alternatively, our framework can be extended to general SUR models by considering priors that impose zero restrictions on certain elements of β.

3. BAYESIAN INFERENCE USING POINT OBSERVATIONS

We now consider the feasibility of a Bayesian analysis of the model in (2.2) – (2.4) on the basis of the recorded point observations, as is the usual practice. Since the prior in (2.3) is improper, we clearly need to verify the existence of the posterior distribution. The following Theorem addresses this issue.

Theorem 1. *Consider n independent replications from (2.2) with any mixing distribution $P_{\lambda_i|\nu}$ and the prior in (2.3) – (2.4) with any proper P_ν. Then the conditional distribution of (β, Σ, ν) given $y \equiv (y_1, \ldots, y_n)'$ exists if and only if $n \geq k + p$.*

Somewhat surprisingly, neither the mixing distribution nor the prior on ν affects the existence of the conditional distribution of the parameters given the observables (*i.e.* the posterior distribution). Thus, whenever $n \geq k+p$, the fact that the prior is improper is of no consequence for the existence of the posterior distribution. However, probability theory tells us that a conditional distribution is only defined up to a set of measure zero in the conditioning variables. In other words, Theorem 1 assures us that $p(y) < \infty$ except possibly on a set of Lebesgue measure zero in $\Re^{n\times p}$. Theoretically, this validates inference since problems can only occur for samples that have zero probability of being observed. However, as stressed in Fernández and Steel (1996a), any recorded sample of point observations has zero probability of occurrence under any continuous sampling distribution. Thus, Theorem 1 does not guarantee that $p(y) < \infty$ for our particular observed sample, and the latter has to be verified explicitly. Note that this problem stems from an inherent violation of the rules of probability calculus, since the recorded observations are in contradiction with the assumed sampling model, and is in no way linked to the improperness of the prior [see Fernández and Steel (1996a) for a more detailed discussion].

If we complement Theorem 1 by considering any possible point $y \in \Re^{n\times p}$, Lemma 1 in the Appendix shows that both $P_{\lambda_i|\nu}$ and P_ν can intervene. It is immediate from Lemma 1 that

$$\text{for finite mixtures of Normals, } p(y) < \infty \text{ if and only if } r(X : y) = k + p, \tag{3.1}$$

which is the minimal possible requirement for any scale mixture of Normals. In the sequel of this Section, we shall, therefore, assume that this rank condition holds.

Let us now analyze the more challenging case of Student-t sampling, where we shall use the following Definition:

Definition 1. *For a design matrix X and a sample $y \in \Re^{n\times p}$, s_j, $j = 1, \ldots, p$, is the largest number of observations such that the rank of the corresponding submatrix of X is k while the rank of the corresponding submatrix of $(X : y)$ is $k + p - j$.*

Clearly, since $r(X : y) = k + p$, we obtain that $k \leq s_p < s_{p-1} < \ldots < s_1 < n$. Now we can present the following Theorem.

Theorem 2. *Let $y = (y_1, \ldots, y_n)'$ be a sample of n independent replications from a p-variate Student-t distribution in* (2.2)*, and consider the prior in* (2.3) − (2.4)*. Assuming that $r(X : y) = k + p$ and defining*

$$m = \max_{j=1,\ldots,p} \left\{ j \frac{n-k}{n-s_j} - p \right\},$$

while recalling Definition 1 for $s_1, \ldots, s_p$, we obtain that
(i) if $m = 0$, then $p(y) < \infty$;
(ii) if $m > 0$, then $p(y) < \infty$ if and only if

$$P_\nu(0, m] = 0 \text{ and } \int_m^{m+\rho} (\nu - m)^{-q} dP_\nu < \infty, \text{ for all } \rho > 0,$$

where q denotes the number of indices $j \in \{1, \ldots, p\}$ for which $m = j\frac{n-k}{n-s_j} - p$.

From Definition 1 we note that $s_j = k + p - j$ (which implies $m = 0$), for all $y \in \Re^{n \times p}$ excluding a set of Lebesgue measure zero. Thus, Theorem 2 (i) will apply and inference is feasible with almost all samples, as was already clear from Theorem 1. However, as will be illustrated in the Examples in Section 4, observed samples often lead to values of $m > 0$, as a consequence of rounding or the finite precision of the measuring device. Then, Theorem 2 (ii) indicates that the prior for ν can not put any mass on values of $\nu \leq m$. As an immediate consequence, inference based on samples for which $m > 0$ is precluded under any prior P_ν with support including $(0, K)$ for some $K > 0$. This negative result even extends to improper priors for ν. Thus, popular choices for P_ν such as the improper Uniform on $\Re_+$, Jeffreys' prior [Liu (1995)] or distributions in the Gamma family [Geweke (1993)] can never lead to a posterior distribution whenever $m > 0$ for the particular sample of point observations under consideration. Bounding ν away from zero by some fixed constant [as in Relles and Rogers (1977) or Liu (1995, 1996)] provides no general solution either, since m is typically updated as sample size grows and can reach an upper bound of $n - k - p$ (when $s_1 = n - 1$). This continual updating of m has the rather shocking consequence that adding new observations can actually destroy the properness of a posterior which was proper with the previous sample!

In the special case of univariate regression ($p = 1$), the quantity m in Theorem 2 simplifies to $m = (s_1 - k)/(n - s_1)$ where s_1 is the largest number of observations such that both the corresponding submatrix of X and the corresponding submatrix of $(X : y)$ have rank k. Now, $s_1 \geq k$ will have the interpretation of the largest possible number of observations for which y_i can be fitted exactly by $\beta' x_i$ for some fixed value of β. Of course, q introduced in Theorem 2 (ii) is one in this case.

If we further specialize to $k = 1$ and take $x_i = 1$, we are in the univariate location-scale model analyzed in Fernández and Steel (1996a). Then, m becomes $(s_1 - 1)/(n - s_1)$, where s_1 is the largest number of observations that are all the same. In that case, as soon as the sample contains repeated observations, a Bayesian analysis on the basis of point observations is precluded if the support of P_ν is not bounded away from zero.

4. BAYESIAN INFERENCE USING SET OBSERVATIONS

A formal solution to the problem mentioned in Section 3 is to consider set observations which have positive probability under the continuous sampling model. In practice, it seems natural to consider a neighbourhood S_i of the recorded point observation y_i on the basis of the precision of the measuring device. This avoids the incompatibility between observations and sampling assumptions and, under a proper prior, posterior inference is always guaranteed. For the improper prior in $(2.3)-(2.4)$ a formal examination leads to the following Theorem:

Theorem 3. *Consider the Bayesian model* $(2.2)-(2.4)$ *and* n *compact sets* S_i, $i=1,\ldots,n$, *of positive Lebesgue measure in* $\Re^p$. *Then* $P(y_1 \in S_1,\ldots,y_n \in S_n) < \infty$ *if and only if* $r(X:y) = k+p$ *for all* $y_1 \in S_1,\ldots,y_n \in S_n$.

Remark that neither the mixing distribution $P_{\lambda_i|\nu}$ nor the prior of ν, P_ν, intervene in Theorem 3, which thus holds for any scale mixture of Normals. Now Bayesian inference is fully coherent and adding extra observations can never destroy the possibility of conducting inference.

The condition $r(X:y) = k+p$, which was always necessary under point observations [see (3.1)], becomes both necessary and sufficient when extended to sets as in Theorem 3. In the case of a location-scale model ($k=1$ and $x_i = 1$), the latter condition is equivalent to the absence of a $(p-1)$-dimensional affine space that intersects with all of the sets $S_1,\ldots,S_n$. Figure 1 graphically illustrates this issue in the bivariate case ($p=2$): while the set observations in Figure 1 (a) allow for posterior inference, the latter is precluded in Figure 1 (b).

In general, Bayesian inference using set observations can easily be implemented through a Gibbs sampler on the parameters augmented with $y = (y_1,\ldots,y_n)'$. We then condition on the set observations $y_1 \in S_1,\ldots,y_n \in S_n$, which shall, for convenience, be denoted as $y \in S$. Let us present this in more detail for Student-t sampling. In this case, it will prove convenient to also augment with the mixing variables $\lambda = (\lambda_1,\ldots,\lambda_n)'$ introduced in (2.2), leading to the following full conditionals:

$$p(y|\beta,\Sigma,\nu,\lambda,y\in S) \propto f_{MN}^{n\times p}(y|X\beta,\Sigma\otimes\Lambda^{-1})I_S(y), \tag{4.1}$$

where we have used the notation for a matricvariate Normal explained in Appendix B, $\Lambda = diag(\lambda_1,\ldots,\lambda_n)$ and $I_S(\cdot)$ denotes the indicator function of the set S. Note that from (4.1) the y_i's are independently drawn from p-variate Normal distributions with mean $\beta' x_i$ and covariance matrix $\lambda_i^{-1}\Sigma$, truncated to the relevant sets S_i.

The second conditional is

$$p(\beta,\Sigma|\nu,y,\lambda,y\in S) = f_{MN}^{k\times p}(\beta|\hat\beta,\Sigma\otimes(X'\Lambda X)^{-1})f_{IW}^{p}(\Sigma|\hat\Sigma,n-k), \tag{4.2}$$

i.e. the product of a matricvariate Normal and an Inverted Wishart density function (as defined in Appendix B), where $\hat\beta = (X'\Lambda X)^{-1}X'\Lambda y$ and $\hat\Sigma = y'\{\Lambda - \Lambda X(X'\Lambda X)^{-1}X'\Lambda\}y$.

The conditional distribution of ν is absolutely continuous with respect to P_ν with Radon-Nikodym derivative proportional to

$$\left(\frac{\nu}{2}\right)^{n\nu/2}\left\{\Gamma\left(\frac{\nu}{2}\right)\right\}^{-n}\exp\left\{-\frac{\nu}{2}\sum_{i=1}^{n}(\lambda_i-\log\lambda_i)\right\}. \tag{4.3}$$

For exponential P_ν (as used in the Examples below), drawings from this non-standard distribution can be generated following Geweke (1994) and Fernández and Steel (1996b).

Finally, n independent Gamma distributions constitute the required conditional for λ:

$$p(\lambda|\beta,\Sigma,\nu,y,y\in S)=\prod_{i=1}^{n}f_G\left(\lambda_i\Big|\frac{\nu+p}{2},\frac{\nu+(y_i-\beta'x_i)'\Sigma^{-1}(y_i-\beta'x_i)}{2}\right), \tag{4.4}$$

where $f_G(\lambda_i|a,b)\propto\lambda_i^{a-1}\exp(-b\lambda_i)$ denotes the probability density function (p.d.f.) of a Gamma distribution.

Using the Gibbs sampler in (4.1) – (4.4), we can easily analyze the following three Examples under Student-t sampling with the prior in (2.3) – (2.4). In all cases, we take P_ν to be exponential with mean 10 and variance 100, *i.e.*

$$p(\nu)=f_G(\nu|1,1/10), \tag{4.5}$$

which spreads the prior mass over a wide variety of tail behaviour. Throughout, results are based on 250,000 Gibbs drawings with a burn-in of 10,000.

We start with a univariate regression model:

Example 1. *Stackloss data*

This classical data set, originally presented in Brownlee (1965), has been subjected to numerous robust methods [*e.g.* Andrews (1974) and Rousseeuw and van Zomeren (1990)] and was treated under Student-t sampling by Lange *et al.* (1989) in a classical maximum likelihood framework. The data consist of $n=21$ observations of a univariate response (stackloss) given an intercept and three other regressors. Thus, $p=1$ and $k=4$. Recalling Definition 1, we can derive that for this data set $s_1=8$, which means that we can fit eight observations y_i by $\beta'x_i$ for a certain value of $\beta\in\Re^4$, namely $\beta=(-36,0.5,1,0)'$. Thus, m in Theorem 2 takes the value $(s_1-k)/(n-s_1)=4/13$, precluding Bayesian inference on the basis of these point observations under Student sampling with the prior in (2.3), (2.4) and (4.5).

Here, we shall conduct inference through set observations in accordance with the precision implicit in the number of digits recorded. We have verified that these set observations fulfill the condition stated in Theorem 3, thus allowing for a Bayesian analysis. Figures 2-5 summarize the posterior inference on the regression coefficients $\beta_2,\ldots,\beta_4$ (excluding the intercept) and the degrees of freedom ν.

The mixing variables λ_i in (2.2) can be seen as observation-specific precision factors, so that unusually small values of λ_i correspond to "outlying" observations. The EM algorithm used in Lange *et al.* (1989) takes the mean of the conditional distribution of λ_i in (4.4) as

the weight of observation i [see also Pettitt (1985) and West (1984)]. On the basis of these weights Lange *et al.* (1989) identify observations 21, 4, 3 and 1 as outliers, like in the least squares analysis of Daniel and Wood (1971) and the robust analysis of Andrews (1974). In a Bayesian setup, we naturally focus on the marginal posterior distribution of the λ_i's and find indeed that the posterior means for these four observations are considerably lower than that for the others. However, we also note that the posterior distributions of the λ_i's display a substantial spread. •

The second example is a multivariate location-scale model:

Example 2. *Fisher's Iris data*

This data set, consisting of $n = 50$ bivariate measurements (of petal length and width) for *Iris setosa*, was analyzed in Fisher (1936) and Heitjan (1989). These data will simply be modelled as a bivariate location-scale model; thus, $p = 2$, $k = 1$ and $x_i = 1$. The original measurements were transformed to logarithms [as suggested in Gnanadesikan (1977) and Heitjan (1989)] and the set observations were transformed accordingly. Heitjan (1989) advocates the use of grouped likelihood for this example and maximizes a Normal likelihood integrated over the respective sets S_i $(i = 1, \ldots, n)$. For these data, we can easily ascertain that $s_1 = 35$ and $s_2 = 29$ (see Definition 1), which implies that $m = 8/3$ and, thus, Theorem 2 (ii) again indicates that point observations can not form the basis of a Bayesian analysis under Student sampling with the prior (2.3), (2.4) and (4.5).

The use of set observations leads to the posterior densities plotted in Figures 6-9, where "correlation" denotes the off-diagonal element of Σ divided by the square root of the product of the diagonal elements. Posterior inference is quite close to Heitjan's classical results. •

A natural extension of our context of set observations is that of missing observations. In a classical analysis of this problem, the EM algorithm was introduced in Dempster, Laird and Rubin (1977), while Bayesian approaches rely on data augmentation [see Tanner and Wong (1987)] or imputation methods [*e.g.* Rubin (1987) and Kong, Liu and Wong (1994)].

Whereas, sofar, we considered observations consisting of bounded sets S_i, the fact that some components of y_i are missing implies that the corresponding set observation becomes unbounded (in the direction of each missing component). Let us now examine the case where $r < n$ observations lead to compact sets, while $n - r$ observations contain unobserved elements.

Theorem 4. *Consider the Bayesian model in* (2.2) − (2.4). *The observations consist of* $r < n$ *compact sets* $S_1, \ldots, S_r$, *whereas* $S_{r+1}, \ldots, S_n$ *are unbounded due to missing components. All* n *sets have positive Lebesgue measure in* $\Re^p$. *Defining* $X_{(r)} = (x_1, \ldots, x_r)'$ *and* $y_{(r)} = (y_1, \ldots, y_r)'$, *we obtain:*
(i) if $r(X_{(r)} : y_{(r)}) = k + p$, *for all* $y_1 \in S_1, \ldots, y_r \in S_r$, *then* $P(y_1 \in S_1, \ldots, y_n \in S_n) < \infty$;
(ii) if we can find values $y_1 \in S_1, \ldots, y_n \in S_n$ *for which* $r(X : y) < k + p$, *then* $P(y_1 \in S_1, \ldots, y_n \in S_n) = \infty$.

From Theorems 3 and 4 (i) we immediately deduce that whenever the compact set observations $S_1, \ldots, S_r$ lead to a proper posterior, the same holds if we add the unbounded set observations $S_{r+1}, \ldots, S_n$ (corresponding to missing data). Clearly, adding observations

that do not contradict the sampling model can never destroy the existence of an already well-defined posterior distribution.

Theorem 4 (ii) addresses the situation where the compact set observations do not result in a posterior [since $r(X_{(r)} : y_{(r)}) < k + p$ for some $y_{(r)}$], and establishes the necessity of $r(X : y) = k + p$ for all $y_1 \in S_1, \ldots, y_n \in S_n$, for conducting Bayesian inference. In other words, Theorem 4 (ii) says that the necessary condition stated in Theorem 3 for compact sets extends to any sample of sets $S_1, \ldots, S_n$ (possibly unbounded).

As explained in the discussion of Theorem 3, the assumption of Theorem 4 (ii) is most easily interpreted in the location-scale case ($k = 1$ and $x_i = 1$), where $r(X : y) < 1 + p$ means that there exists a $(p-1)$-dimensional affine space that intersects all of the sets $S_1, \ldots, S_n$. Note that if we can find one such space that, in addition, intersects with all p coordinate axes, any new set corresponding to missing data will necessarily have an intersection with this $(p-1)$-dimensional affine space. Thus, adding any number of sets corresponding to observations with missing components can never result in a posterior distribution. Figure 10 graphically illustrates this point in the bivariate case ($p = 2$).

Example 3. *Extended Murray data*

In this Example, we focus on the artificial data originally introduced by Murray (1977) and extended with four extreme values in Liu and Rubin (1995). This results in the following bivariate data set:

$$\begin{array}{cccccccccccccccc} y_{i1} & -1 & -1 & 1 & 1 & -2 & -2 & 2 & 2 & ? & ? & ? & ? & -12 & 12 & ? & ? \\ y_{i2} & -1 & 1 & -1 & 1 & ? & ? & ? & ? & -2 & -2 & 2 & 2 & ? & ? & -12 & 12 \end{array} \tag{4.6}$$

where $n = 16$, $p = 2$ and ? denotes a missing value. We use a location-scale model under Student-t sampling, as in the maximum likelihood analysis of Liu and Rubin (1995) and the Bayesian analysis of Liu (1995). Here, Bayesian inference will be conducted under the prior (2.3), (2.4) and (4.5), using set observations with a unitary width for the observed components.

Figure 11 depicts the four compact sets (corresponding to the first four observations), and from Theorem 4 (i) we can immediately deduce properness of the posterior as no single line can cross all four sets $S_1, \ldots, S_4$. Note that deleting any one of these compact set observations would result in an improper posterior [see Theorem 4 (ii) and the discussion thereafter] as the dashed line in Figure 11 indicates. Figures 12-13 plot posterior p.d.f.'s for the correlation (as defined in Example 2) and the degrees of freedom ν. Compared to the same analysis on the basis of the original Murray data [*i.e.* the first twelve in (4.6)], the correlation is more extreme and degrees of freedom tend to be substantially smaller. As expected, the analysis based on all sixteen set observations identifies the four extra observations [*i.e.* the last four in (4.6)] as outliers through small values of the mixing variables λ_i associated with these observations. In addition, we have found that inference with this model is remarkably insensitive with respect to the width chosen for the set observations (in directions where they are bounded). •

5. THE STUDENT-t LIKELIHOOD FUNCTION

In this Section, we examine some peculiarities of the likelihood function corresponding to independent Student-t sampling in a general regression context. We shall only focus on the use of point observations, and present some classical counterparts of the problems described in Section 3 under a Bayesian treatment of this model.

In particular, we consider n replications from the following sampling density function for $y_i \in \Re^p$:

$$p(y_i|\beta,\Sigma,\nu) = \frac{\Gamma\{(\nu+p)/2\}}{\Gamma(\nu/2)(\pi\nu)^{p/2}|\Sigma|^{1/2}}\left[1+\frac{1}{\nu}\{y_i-g_i(\beta)\}'\Sigma^{-1}\{y_i-g_i(\beta)\}\right]^{-(\nu+p)/2}, \quad (5.1)$$

where $\beta \in \mathcal{B}$ and $g_i(\cdot)$ is a known continuous function from $\mathcal{B}$ to $\Re^p$, possibly depending on regressors x_i. Thus, we extend the linear regression context of the previous Sections to more general regression functions.

We shall reparameterize the PDS matrix Σ as (σ, V) through

$$\Sigma = \sigma^2 V, \quad (5.2)$$

where $\sigma \in \Re_+$ and $V \in \mathcal{C}_1^p$, which will denote the space of $p \times p$ PDS matrices with element $(1,1)$ equal to one. This reparameterization is useful for presenting the main result of this Section.

Theorem 5. *Let $l(\beta,\sigma,V,\nu)$ be the likelihood function corresponding to n independent replications from* (5.1) *with the reparameterization in* (5.2)*. Then:*
(i) $l(\beta,\sigma,V,\nu)$ is a finite continuous function in the entire parameter space $\mathcal{B}\times(0,\infty)\times\mathcal{C}_1^p\times(0,\infty)$.
(ii) For given values $\beta=\beta_0$, $V=V_0$ and $\nu=\nu_0$, let $0 \le s(\beta_0) \le n$ be the number of observations for which $y_i = g_i(\beta_0)$. We obtain:
(iia)

$$\text{if } \nu_0 < \frac{s(\beta_0)p}{n-s(\beta_0)}, \text{ then } \lim_{\sigma\to 0} l(\beta_0,\sigma,V_0,\nu_0) = \infty;$$

(iib)

$$\text{if } \nu_0 = \frac{s(\beta_0)p}{n-s(\beta_0)}, \text{ then } \lim_{\sigma\to 0} l(\beta_0,\sigma,V_0,\nu_0) \in (0,\infty);$$

(iic)

$$\text{if } \nu_0 > \frac{s(\beta_0)p}{n-s(\beta_0)}, \text{ then } \lim_{\sigma\to 0} l(\beta_0,\sigma,V_0,\nu_0) = 0.$$

From Theorem 5, whenever we can find a value β_0 such that $y_i = g_i(\beta_0)$ holds for at least one observation, the likelihood function does not possess a global maximum. Indeed, for small enough values of ν [see (iia)], we can make $l(\beta_0,\sigma,V_0,\nu_0)$ arbitrarily large by letting σ tend to zero. Note that, in practice, we can typically find values β_0 such

that $s(\beta_0) > 0$. For example, in the case of p-variate linear regression with k regressors (considered in the previous Sections), we can deduce from Definition 1 that

$$\max_{\beta \in \Re^{k \times p}} s(\beta) = s_p \geq k, \tag{5.3}$$

thus precluding global maximization of the likelihood function, irrespectively of the sample.

This finding casts some doubt on maximum likelihood (ML) estimation under Student-t regression models with unknown degrees of freedom ν. In the existing literature, ν is typically allowed to vary in $\Re_+$ [see *e.g.* Lange *et al.* (1989) and Lange and Sinsheimer (1993)]. Reported ML estimates must, therefore, correspond to local and not to global maxima, although this is not stated in these papers. To our knowledge, the existence and uniqueness of such local maxima and the asymptotic properties of the corresponding estimators have not been formally established in the literature. In a pure location-scale context with fixed degrees of freedom, constrained to be sufficiently large, Maronna (1976) proves that the likelihood equations have a unique solution that leads to a consistent and asymptotically Normal estimator of (β, Σ). However, we have not encountered similar results for unknown ν.

We remind the reader that a Bayesian analysis of the linear regression model based on point observations breaks down if we assign prior probability to values of $\nu \leq m$, where m is defined in Theorem 2. From Theorem 5 (ii) with (5.3) the likelihood is unbounded if $\nu < s_p p/(n - s_p)$, which can generally be larger or smaller than m. In the case of univariate linear regression ($p = 1$), the latter quantity becomes $s_1/(n - s_1)$, which is always larger than $m = (s_1 - k)/(n - s_1)$. Thus, in this case, the likelihood is still integrable with the prior in $(2.3) - (2.4)$ if P_ν bounds ν strictly away from $(s_1 - k)/(n - s_1)$ (Theorem 2), but is unbounded for values of ν smaller than $s_1/(n - s_1)$. Furthermore, there is a fundamental difference between classical and Bayesian results: as remarked in the discussion of Theorem 2, m equals zero for all $y \in \Re^{n \times p}$ except for a set of Lebesgue measure zero, implying that a Bayesian analysis is feasible for almost all samples (see also Theorem 1). In practice problems only occur due to the coarse nature of observed data. From Theorem 5 (iia) and (5.3), on the other hand, it is immediately clear that global maximization of the likelihood is precluded for any sample $y \in \Re^{n \times p}$.

In order to illustrate the behaviour of the likelihood function, let us reconsider Example 1.

Example 1. (continued) *Stackloss data*

As explained in Example 1, the value $\beta_0 = (-36, 0.5, 1, 0)'$ allows us to exactly fit eight of the 21 observations. Thus, from Theorem 5 (iia), taking $\nu_0 < 8/13$ leads to $\lim_{\sigma \to 0} l(\beta_0, \sigma, 1, \nu_0) = \infty$ (note that $p = 1$ implies $V_0 = 1$).

Figure 14 plots the logarithm of $l(\beta_0, \sigma, 1, \nu_0)$ as a function of the logarithm of σ, for β_0 as above and different values of ν_0, illustrating Theorem 5 (ii). Values of ν_0 smaller than $8/13$ clearly lead to an unbounded likelihood, for $\nu_0 = 8/13$ the likelihood converges to a positive finite value as $\sigma \to 0$, whereas $\nu_0 > 8/13$ leads to a zero limit as σ tends to zero. From the form of the likelihood function it is immediate that for small values of σ, the log likelihood is approximately linear in $ln(\sigma)$ with slope coefficient $\nu_0\{n - s(\beta_0)\} - s(\beta_0)p$, which is also apparent from Figure 14.

Lange *et al.* (1989) estimate ν to be 1.1, which presumably corresponds to a local maximum of the likelihood. •

In some cases, numerical optimization procedures (such as the EM algorithm) may attempt to converge to an area with unbounded likelihood. A case in point is the analysis of the radioimmunoassay data in Lange *et al.* (1989) and Lange and Sinsheimer (1993). This concerns a nonlinear regression model with $p = 1$ (*i.e.* univariate) and 4 regression parameters introduced in Tiede and Pagano (1979), where the $n = 14$ data points are listed. Whereas Lange *et al.* (1989, p.883) already report that "ML estimation of ν for this data is not very satisfactory" and report an ML estimate of ν equal to 0.29, Lange and Sinsheimer (1993) report ML estimates of ν equal to 0.05 and of σ equal to 0. The latter also state that 10 of the 14 weights [*i.e.* the mean of λ_i in (4.4)] are found to be zero and the EM algorithm has not converged after 300 iterations. From their estimates of β it is clear that they exactly fit four of the observations, while they consider values of ν smaller than 4/10, which takes them to a region of unbounded likelihood. Thus, Theorem 5 provides an immediate explanation for the "potential problems with the t" mentioned in Lange and Sinsheimer (1993, p.195).

When some of the components of the p-variate observations y_i, $i = 1, \ldots, n$ are missing, the resulting likelihood still displays the same type of behaviour as explained in Theorem 5. In particular, the latter Theorem will apply in this more general context if we replace the bound $\{s(\beta_0)p\}/\{n - s(\beta_0)\}$ for ν by the quantity

$$\frac{\sum_{i \in \mathcal{I}} p_i}{n - s(\beta_0)},$$

where $p_i \leq p$ is the number of observed components of y_i, $\mathcal{I}$ is the set of indices for which the observed components of y_i are exactly fitted by the corresponding components of $g_i(\beta_0)$, and $s(\beta_0)$ is the cardinality of $\mathcal{I}$. Thus, the stationary values reported in Liu and Rubin (1994, 1995) for Student-t models with unknown ν and missing data do not correspond to global maxima of the likelihood function.

In conclusion, we feel that the use of ML methods for Student-t models with unknown degrees of freedom can not be advocated without further careful study of the existence and properties of local maxima. Alternatively, classical inference could be based on efficient likelihood estimation [Lehmann (1983, chap. 6)], grouped likelihoods [see Giesbrecht and Kempthorne (1976) for a lognormal model and Beckman and Johnson (1987) for the Student-t case], sample percentiles [Resek (1976)], modified likelihood [as in Cheng and Iles (1987)] or spacings methods [as in Cheng and Amin (1979)]. For a general discussion of non-regular likelihood problems, see Smith (1989) and Cheng and Traylor (1995).

6. CONCLUSION

In this paper we considered likelihood-based inference from multivariate regression models with errors that are distributed as scale mixtures of Normals. Some very intruiging pitfalls of both Bayesian and classical methods are uncovered. A fully coherent procedure is proposed from a Bayesian point of view.

The Bayesian model consists of independent sampling from a linear regression model with a scale mixture of Normals error distribution, combined with a commonly used improper reference prior. Usually, Bayesian analysis is conducted given a sample of point observations. We show (Theorem 1) that the conditional distribution of the parameters θ given the observables y exists if and only if sample size $n \geq k + p$, where k is the number of regressors and p is the dimension of the response variable. Thus, it seems that the extension of the sampling distribution from multivariate Normal to the entire class of scale mixtures of multivariate Normals leaves the existence of the posterior entirely unaffected. There are, however, two crucial facts to be noted: firstly, a conditional distribution is defined up to a set of measure zero in the conditioning variable, and, secondly, any sample y_0 of point observations that we record has probability zero of occurring under a continuous sampling model. Thus, Theorem 1 does not assure us that $p(y_0) < \infty$ for our particular sample, and this needs to be verified explicitly. It turns out that the set of measure zero for which $p(y) = \infty$ does depend on the mixing distribution. For the leading case of Student-t sampling with unknown degrees of freedom, ν, Theorem 2 characterizes the samples for which $p(y) < \infty$. Many samples that are likely to occur on practice (due to rounding or finite precision) are seen to require a positive lower bound on ν. As this lower bound changes with each new observation and can get as large as $n - k - p$, the usual Bayesian analysis given point observations can not be recommended as a generally applicable procedure. Once a posterior is found to exist, this does not guarantee inference on the basis of an extended sample!

This problem, which derives from a fundamental incompatibility between the continuous sampling model and point observations [see Fernández and Steel (1996a) for a more detailed discussion], is solved by considering set observations. Instead of on the actually recorded value, we condition inference on a set around each recorded value. The necessary and sufficient condition that validates Bayesian inference using set observations is exactly the same for every member in the class of scale mixtures of Normals (Theorem 3). All we need is that the full column rank condition on the matrix of regressors and observables, $(X : y)$, holds for all values of y in the set observations we consider. The analysis is now fully coherent, in that new observations can never destroy the possibility of conducting inference.

A simple Gibbs sampling strategy is proposed to implement Bayesian analysis with set observations and a number of Examples is considered, all under Student sampling with an Exponential prior on ν. A univariate regression model with $k = 4$ regressors is used for the well-known stackloss data [Brownlee (1965)], whereas the Fisher iris data are handled with a bivariate location-scale model. In both cases, we find that a Bayesian analysis using point observations is precluded if ν is not bounded away from zero. Bayesian inference through set observations is, however, quite feasible and requires only moderate numerical effort. The identification of outliers is straightforward.

We also consider the case where some components of the p-variate response are not observed, *i.e.* missing data. It is seen in Theorem 4 that observations with missing components will typically not help in establishing properness of the posterior distribution. The artificial data set used in Liu and Rubin (1995) is analyzed in Example 3.

A closer look at the likelihood function of a multivariate Student-t model with possibly

nonlinear regression leads to the following finding: Using point observations, the likelihood function is unbounded as we tend to the boundary of the parameter space for small enough values of ν (Theorem 5). This result, which also generalizes to the case with missing data, raises questions regarding the interpretation and validity of maximum likelihood for Student models with unknown ν. This immediately provides an explanation for problems such as encountered by Lange *et al.* (1989) and Lange and Sinsheimer (1993) in the analysis of the radioimmunoassay data from Tiede and Pagano (1979). Even if local maxima are found with numerical techniques, such as the EM algorithm, the theoretical properties of the corresponding estimators seem, as yet, not established in the literature.

Although this behaviour of the likelihood function is, of course, related to the Bayesian results on existence of the posterior based on point observations, there are some important differences. The restrictions on ν required to avoid the problems in the univariate regression model ($p = 1$) are stronger for classical inference than for Bayesian inference. More importantly, whereas Bayesian inference is only precluded for a set of Lebesgue measure zero in the observables (and problems may occur in practice due to rounding), the likelihood will always be unbounded, irrespectively of the sample.

In summary, extending the error distribution of regression models to independent Student-t sampling is not as innocuous an extension from a theoretical point of view as it might seem from a merely numerical angle. Whereas computational methods to analyze such models are readily available [Monte Carlo Markov Chain methods such as Gibbs sampling for Bayesian inference, and EM-type algorithms for ML], they should not be applied blindly, since some theoretical pitfalls may preclude inference in actual practice.

APPENDIX A: PROOFS OF THEOREMS

Throughout the Appendices, $|\cdot|$ will stand for determinant, $tr(\cdot)$ will denote trace and $\mathcal{C}^p$ the set of $p \times p$ PDS matrices. The following Lemma will be instrumental for proving the Theorems:

Lemma 1. *Consider $y = (y_1, \ldots, y_n)' \in \Re^{n \times p}$ a sample of n independent replications from (2.2) and the prior in (2.3) − (2.4). Then $p(y) < \infty$ if and only if $r(X : y) = k + p$ and*

$$\int_{(0,\infty)^n} \Big(\prod_{i \neq m_1, \ldots, m_{k+p}} \lambda_i^{p/2} \Big) \Big(\prod_{i=k+1}^{k+p} \lambda_{m_i}^{-(n-k-p)/2} \Big) dP_{(\lambda_1,\ldots,\lambda_n)} < \infty, \qquad (A.1)$$

where we have defined

$$P_{(\lambda_1,\ldots,\lambda_n)} \equiv \int_{\mathcal{N}} \left(\prod_{i=1}^{n} P_{\lambda_i|\nu} \right) dP_\nu \ \ (\textit{with a slight abuse of notation}), \qquad (A.2)$$

$$\prod_{i=1}^{k} \lambda_{m_i} \equiv \max \left\{ \prod_{i=1}^{k} \lambda_{l_i} : |x_{l_1} \ldots x_{l_k}| \neq 0 \right\} \qquad (A.3)$$

and

$$\prod_{i=1}^{k+p} \lambda_{m_i} \equiv \max \left\{ \prod_{i=1}^{k+p} \lambda_{l_i} : \begin{vmatrix} x_{l_1} & \dots & x_{l_{k+p}} \\ y_{l_1} & \dots & y_{l_{k+p}} \end{vmatrix} \neq 0 \right\}. \tag{A.4}$$

Proof: A sample of n independent replications from (2.2) with the prior in (2.3) – (2.4) leads to

$$p(y) \propto \int_{\Re^{k\times p}\times \mathcal{C}^p \times (0,\infty)^n} \frac{|\Lambda|^{p/2}}{|\Sigma|^{(n+p+1)/2}} \exp\left[-\frac{tr\{\Sigma^{-1}(y-X\beta)'\Lambda(y-X\beta)\}}{2} \right] d\beta d\Sigma dP_{(\lambda_1,\dots,\lambda_n)}, \tag{A.5}$$

where $\Lambda = diag(\lambda_1, \dots, \lambda_n)$ and $P_{(\lambda_1,\dots,\lambda_n)}$ was defined in (A.2). The integrand in (A.5) is proportional to

$$\frac{|\Lambda|^{p/2}}{|X'\Lambda X|^{p/2}} |\Sigma|^{-(n-k+p+1)/2} \exp\left\{ -\frac{tr(\Sigma^{-1}\hat{\Sigma})}{2} \right\} f_{MN}^{k\times p}(\beta|\hat{\beta}, \Sigma \otimes (X'\Lambda X)^{-1}), \tag{A.6}$$

where $\hat{\beta} = (X'\Lambda X)^{-1}X'\Lambda y$, $\hat{\Sigma} = y'\{\Lambda - \Lambda X(X'\Lambda X)^{-1}X'\Lambda\}y$ and we use the notation for the matricvariate Normal density function introduced in Appendix B. From the expression in (A.6) we note that β can immediately be integrated out, whereas standard distribution theory results show that integrating out Σ requires $n \geq k+p$ and $\hat{\Sigma}$ to be a PDS matrix. Thus, we need to impose that $|\hat{\Sigma}| > 0$. It is easy to see that

$$|\hat{\Sigma}| = \frac{|L'L|}{|X'\Lambda X|}, \text{ with } L = \Lambda^{1/2}(X : y). \tag{A.7}$$

Since $r(X) = k$, $|X'\Lambda X| > 0$; therefore, $|\hat{\Sigma}| > 0$ if and only if $r(L) = k+p$, which is equivalent to $r(X : y) = k+p$. In summary, we have obtained that

$$p(y) < \infty \text{ requires } r(X : y) = k+p.$$

When this rank condition holds, we can integrate out Σ using an Inverted Wishart distribution, which leaves us with [see (A.6), (A.7) and (B.2)] a constant times

$$\frac{|\Lambda|^{p/2}}{|X'\Lambda X|^{p/2}|\hat{\Sigma}|^{(n-k)/2}} = \frac{|\Lambda|^{p/2}|X'\Lambda X|^{(n-k-p)/2}}{|L'L|^{(n-k)/2}} \tag{A.8}$$

to be integrated with respect to $P_{(\lambda_1,\dots,\lambda_n)}$. Applying the Binet-Cauchy formula [see Gantmacher (1959, p. 9)] leads to

$$\begin{aligned} |X'\Lambda X| &= \sum_{1\leq l_1<\dots<l_k\leq n} \left(\prod_{i=1}^{k} \lambda_{l_i} \right) |x_{l_1} \dots x_{l_k}|^2 \\ \text{and } |L'L| &= \sum_{1\leq l_1<\dots<l_{k+p}\leq n} \left(\prod_{i=1}^{k+p} \lambda_{l_i} \right) \begin{vmatrix} x_{l_1} & \dots & x_{l_{k+p}} \\ y_{l_1} & \dots & y_{l_{k+p}} \end{vmatrix}^2. \end{aligned} \tag{A.9}$$

Thus, $|X'\Lambda X|$ has upper and lower bounds both proportional to $\prod_{i=1}^{k}\lambda_{m_i}$, defined in $(A.3)$, whereas $|L'L|$ has upper and lower bounds both proportional to $\prod_{i=1}^{k+p}\lambda_{m_i}$, defined in $(A.4)$. This implies that integrability of the expression in $(A.8)$ with respect to $P_{(\lambda_1,\ldots,\lambda_n)}$ is equivalent to $(A.1)$, thus concluding the proof of the Lemma.

Proof of Theorem 1

The conditional distribution of (β,Σ,ν) exists if and only if $p(y)<\infty$ for all $y\in\Re^{n\times p}$, possibly excluding a set of Lebesgue measure zero.

From Lemma 1, it is immediate that $p(y)<\infty$ always requires $n\geq k+p$.

On the other hand, if $n\geq k+p$, then for all $y\in\Re^{n\times p}$ excluding a set of Lebesgue measure zero, $r(X:y)=k+p$ and $\max\{\lambda_i : i\neq m_1,\ldots,m_{k+p}\}\leq\min\{\lambda_{m_i} : i=k+1,\ldots,k+p\}$ [see $(A.3)$ and $(A.4)$]. This implies a bounded integrand in $(A.1)$, which, in turn, implies that $(A.1)$ holds since $P_{(\lambda_1,\ldots,\lambda_n)}$ is a probability distribution. Sufficiency of $n\geq k+p$ is now immediate from Lemma 1.

Proof of Theorem 2

In order to check whether $(A.1)$ holds, we shall decompose the domain of integration, $(0,\infty)^n$, into all $n!$ possible orderings of $\lambda_1,\ldots,\lambda_n$. It is enough to focus on those orderings for which $\lambda_{(n-s_j:n)}=\lambda_{m_{k+j}}$ for all $j=1,\ldots,p$ [where s_j was given in Definition 1 and $\lambda_{(i:n)}$ denotes the i^{th} order statistic] since they lead to the largest value for the integrand in $(A.1)$. For any such ordering, we first integrate with respect to $\prod_{i=1}^{n}P_{\lambda_i|\nu}$ [where $P_{\lambda_i|\nu}$ is a Gamma$(\nu/2,\nu/2)$ distribution] and finally with respect to P_ν. In each of the steps of the integration process we shall use the upper and lower bounds

$$\exp(-b\lambda_{i+1})\frac{\lambda_{i+1}^a}{a}\leq\int_0^{\lambda_{i+1}}\lambda_i^{a-1}\exp(-b\lambda_i)d\lambda_i\leq\frac{\lambda_{i+1}^a}{a},\ \text{for any } a,b>0. \tag{A.10}$$

Iterative use of the lower bound in $(A.10)$ directly shows that a finite integral in $(A.1)$ requires $P_\nu(0,m]=0$, with m as defined in Theorem 2.

If we now assume that $P_\nu(0,m]=0$, we can integrate with respect to $\prod_{i=1}^{n}P_{\lambda_i|\nu}$ and, applying $(A.10)$, we obtain a lower bound proportional to

$$h(\nu)n^{-n\nu/2}\nu^{1-s_p}, \tag{A.11}$$

and an upper bound proportional to

$$h(\nu)\{(n-s_p+1)\nu-p(s_p-k)\}^{1-s_p}, \tag{A.12}$$

where we have defined

$$h(\nu)\equiv\frac{\Gamma(n\nu/2)}{\Gamma(\nu/2)^n}(\nu+p)^{-(n-s_1-1)}\left[\prod_{j=1}^{p-1}\prod_{l=n-s_j}^{n-s_{j+1}-1}\left\{\nu-\left(j\frac{n-k}{l}-p\right)\right\}^{-1}\right]\left(\nu-p\frac{s_p-k}{n-s_p}\right)^{-1}. \tag{A.13}$$

(i) $m=0$ clearly implies $s_p=k$, and the upper bound in $(A.12)$ is finite for all $\nu>0$ and has a finite limit as ν converges to zero. Thus, integrability over any finite interval $(0,K)$ is assured under any proper prior P_ν.

(ii) If $m > 0$, the integral condition stated in Theorem 2 (ii) is clearly necessary for integrability of $(A.11)$, whereas it also guarantees integrability of $(A.12)$ over any finite interval (m, K).

Finally, we need to examine integrability over an unbounded interval (K, ∞). We shall consider the following upper bound for the integrand in $(A.1)$:

$$\frac{\lambda_m^{p(n-k-p)/2}}{\lambda_{m_{k+1}}^{p(n-k-p)/2}}, \text{ where } \lambda_m = \max\{\lambda_i : i \neq m_1, \ldots, m_{k+p}\}, \lambda_{m_{k+1}} = \min\{\lambda_{m_{k+1}}, \ldots, \lambda_{m_{k+p}}\}.$$

Thus, the integral with respect to $\prod_{i=1}^n P_{\lambda_i|\nu}$ is bounded from above by

$$\int_0^\infty \lambda_m^{p(n-k-p)/2} f_G\left(\lambda_m|\frac{\nu}{2}, \frac{\nu}{2}\right) d\lambda_m \int_0^\infty \lambda_{m_{k+1}}^{-p(n-k-p)/2} f_G\left(\lambda_{m_{k+1}}|\frac{\nu}{2}, \frac{\nu}{2}\right) d\lambda_{m_{k+1}}$$
$$= \Gamma\left\{\frac{\nu + p(n-k-p)}{2}\right\} \Gamma\left\{\frac{\nu - p(n-k-p)}{2}\right\} \Gamma\left(\frac{\nu}{2}\right)^{-2}.$$

The latter function of ν takes finite values if $\nu \geq K > p(n-k-p)$ and has a finite limit as ν tends to ∞. This leads to a finite integral over the range (K, ∞) under any proper prior P_ν.

Proof of Theorem 3

After integrating out β and Σ as in the proof of Lemma 1, we need to integrate what remains from the expression in $(A.6)$ over $y_1 \in S_1, \ldots, y_n \in S_n$ and, finally, with respect to $P_{(\lambda_1,\ldots,\lambda_n)}$. Since $r(X) = k$ we assume, without loss of generality, that $|x_1 \ldots x_k| \neq 0$. Defining

$$\eta = \begin{pmatrix} \eta_1' \\ \ldots \\ \eta_n' \end{pmatrix} \equiv y - X \begin{pmatrix} x_1' \\ \ldots \\ x_k' \end{pmatrix}^{-1} \begin{pmatrix} y_1' \\ \ldots \\ y_k' \end{pmatrix}, \tag{A.13}$$

which clearly implies $\eta_1 = \ldots = \eta_k = 0 \in \Re^p$, leads to the following expression for $\hat{\Sigma}$ [defined after $(A.6)$]:

$$\begin{aligned} &\hat{\Sigma} = \eta_{-k}' Q(\Lambda) \eta_{-k}, \text{ where} \\ &\eta_{-k} = (\eta_{k+1}, \ldots, \eta_n)' \in \Re^{(n-k)\times p} \text{ and } Q(\Lambda) = \Lambda_{-k} - \Lambda_{-k} X_{-k} (X'\Lambda X)^{-1} X_{-k}' \Lambda_{-k}, \\ &\text{with } \Lambda_{-k} = diag(\lambda_{k+1}, \ldots, \lambda_n) \text{ and } X_{-k} = (x_{k+1}, \ldots, x_n)'. \end{aligned} \tag{A.14}$$

Thus, we are left with

$$|\Lambda|^{p/2} |X'\Lambda X|^{-p/2} |\eta_{-k}' Q(\Lambda) \eta_{-k}|^{-(n-k)/2}, \tag{A.15}$$

which we need to integrate in $y_1, \ldots, y_k, \eta_{k+1}, \ldots, \eta_n$ over the appropriate sets [note that this change of variables has unitary Jacobian], and with respect to $P_{(\lambda_1,\ldots,\lambda_n)}$.

Necessity of $r(X:y) = k+p$ for all $y_1 \in S_1, \ldots, y_n \in S_n$:

Let us assume that there exist $y_1 \in S_1, \ldots, y_n \in S_n$ for which $r(X:y) < k+p$. From $(A.7)$ and $(A.14)$, this condition is equivalent to

$$|\eta'_{-k} Q(\Lambda) \eta_{-k}| = 0, \text{ for some } y_1 \in S_1, \ldots, y_n \in S_n. \tag{A.16}$$

In order to integrate out η_{-k} from $(A.15)$, we subsequently consider the following changes of variables

$$\begin{aligned} &\eta_{-k} \in \Re^{(n-k)\times p} \longrightarrow \xi = Q(\Lambda)^{1/2} \eta_{-k} \in \Re^{(n-k)\times p} \\ &\longrightarrow \theta = \xi'\xi \in \mathcal{C}^p \text{ (completed with } (n-k)p - \{p(p+1)/2\} \text{ extra variables)} \\ &\longrightarrow T, \text{ an upper triangular matrix of order } p : \theta = T'T. \end{aligned} \tag{A.17}$$

This transformation has Jacobian

$$|Q(\Lambda)|^{-p/2} \left(\prod_{j=1}^{p} t_{11}^2 \ldots t_{jj}^2 \left| \begin{matrix} \xi_{11} & \ldots & \xi_{1j} \\ \ldots & \ldots & \ldots \\ \xi_{j1} & \ldots & \xi_{jj} \end{matrix} \right|^{-2} \right)^{1/2} \geq |Q(\Lambda)|^{-p/2}, \tag{A.18}$$

where t_{ij} and ξ_{ij} respectively denote the $(i,j)^{th}$ elements of T and ξ, and where the inequality can be proven by means of the Binet-Cauchy formula. From $(A.17)$ and $(A.18)$, we see that integrating out η_{-k} from $(A.15)$ requires that the integral

$$\int \prod_{j=1}^{p} t_{jj}^{-(n-k)} dt_{11} \ldots dt_{pp}$$

is finite. This, however, does not hold since, by the assumption in $(A.16)$, there is a point in the domain of integration where $\prod_{j=1}^{p} t_{jj} = 0$.

Sufficiency of $r(X:y) = k+p$ for all $y_1 \in S_1, \ldots, y_n \in S_n$:

From the definition of η in $(A.13)$, this condition is equivalent to $r(X:\eta) = k+p$, which is, in turn, equivalent to $|\eta'_{-k}\eta_{-k}| > 0$ for all $y_1 \in S_1, \ldots, y_n \in S_n$. Since we have assumed that these sets are compact, this implies $|\eta'_{-k}\eta_{-k}| \geq A > 0$ for some positive constant A and, by the Binet-Cauchy formula, this means that there always exists a submatrix of η_{-k} of order p with determinant strictly bounded away from zero. Let us *e.g.* consider the region where

$$|\eta_{k+1} \ldots \eta_{k+p}|^2 \geq B > 0 \text{ for some constant } B. \tag{A.19}$$

Direct use of linear algebra shows that $(A.15)$ is proportional to

$$\frac{|\Lambda|^{p/2}}{|X'\Lambda X|^{p/2} |Q(\Lambda)|^{p/2}} (|\eta_{k+1} \ldots \eta_{k+p}|^2)^{-p/2} f_{MS}^{(n-k-p)\times p}(\eta_{-(k+p)} | \hat{\eta}, R, P, p), \tag{A.20}$$

where we use the notation for the matricvariate Student distribution introduced in Appendix B, and we have defined $\eta_{-(k+p)} = (\eta_{k+p+1}, \ldots, \eta_n)' \in \Re^{(n-k-p)\times p}$,

$$Q(\Lambda) = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \text{ with } Q_{11} \text{ of order } p \times p, \tag{A.21}$$

$h = (\eta_{k+1}, \ldots, \eta_{k+p})'$, $\hat{\eta} = -Q_{22}^{-1}Q_{21}h$, $R = h'(Q_{11} - Q_{12}Q_{22}^{-1}Q_{21})h$ and $P = Q_{22}$. From the definitions of $Q(\Lambda)$ in $(A.14)$ and Q_{22} in $(A.21)$, the latter matrix is clearly PDS, whereas the assumption in $(A.19)$ implies that R is also a PDS matrix; thus, we can integrate out $\eta_{-(k+p)}$ in $(A.20)$ using a proper matricvariate Student distribution. From $(A.19)$ and the fact that we have a bounded domain of integration, we can also integrate out $\eta_{k+1}, \ldots, \eta_{k+p}, y_1, \ldots, y_k$, obtaining a finite integral. Finally, from the definition of $Q(\Lambda)$ in $(A.14)$ we can derive that $|Q(\Lambda)| \propto |\Lambda||X'\Lambda X|^{-1}$, which, in combination with $(A.20)$, leads to a constant integrand, and thus a finite integral with respect to the probability distribution $P_{(\lambda_1,\ldots,\lambda_n)}$.

Proof of Theorem 4

(i) By Theorem 3, the assumption of Theorem 4 (i) implies that $P(y_1 \in S_1, \ldots, y_r \in S_r) < \infty$. Straightforward calculations show that, since $r \leq n$, $P(y_1 \in S_1, \ldots, y_n \in S_n) \leq P(y_1 \in S_1, \ldots, y_r \in S_r) < \infty$.

(ii) Immediate, since the proof of the necessity in Theorem 3 never uses the fact that the sets are compact.

Proof of Theorem 5

The result follows immediately from writing down the likelihood function.

APPENDIX B: MATRICVARIATE DENSITY FUNCTIONS

Here we present the density functions of the matricvariate distributions used in the paper.

Matricvariate Normal:

The $p \times q$ random matrix A has a matricvariate Normal distribution with mean $M \in \Re^{p\times q}$ and covariance matrix of the column expansion $vec(A)$ given by $\Omega \otimes P$, where $\Omega \in \mathcal{C}^q$ and $P \in \mathcal{C}^p$, if the density function of A is:

$$f_{MN}^{p\times q}(A|M, \Omega \otimes P) \equiv \{(2\pi)^{pq}|\Omega|^p|P|^q\}^{-1/2} \exp\left[-\frac{tr\{\Omega^{-1}(A-M)'P^{-1}(A-M)\}}{2}\right]. \quad (B.1)$$

Inverted Wishart:

The random matrix $S \in \mathcal{C}^q$ has an Inverted Wishart distribution if its density function is given by:

$$f_{IW}^{q\times q}(S|Q, \nu) \equiv \left\{2^{\nu q/2}\pi^{q(q-1)/4}\prod_{i=1}^{q}\Gamma\left(\frac{\nu+1-i}{2}\right)\right\}^{-1} |Q|^{\nu/2}|S|^{-(\nu+q+1)/2} \exp\left\{-\frac{tr(S^{-1}Q)}{2}\right\}, \quad (B.2)$$

where $Q \in \mathcal{C}^q$ and $\nu > q - 1$.

Matricvariate Student:

The $p \times q$ random matrix A has a matricvariate Student-t distribution if it has the following density function:

$$f_{MS}^{p\times q}(A|M,Q,H,\nu) \equiv \pi^{-pq/2} \prod_{i=1}^{q} \left[\Gamma\left(\frac{\nu+p+1-i}{2}\right) \left\{ \Gamma\left(\frac{\nu+1-i}{2}\right) \right\}^{-1} \right] \times |Q|^{\nu/2} |H|^{q/2} |Q + (A-M)'H(A-M)|^{-(\nu+p)/2}, \tag{B.3}$$

where $M \in \Re^{p\times q}$, $Q \in \mathcal{C}^q$, $H \in \mathcal{C}^p$ and $\nu > q-1$.

REFERENCES

Andrews, D.F. (1974), "A Robust Method for Multiple Linear Regression," *Technometrics*, 16, 523-531.

Beckman, R.J., and Johnson, M.E. (1987), "Fitting the Student-t Distribution to Grouped Data, With Application to a Particle Scattering Experiment," *Technometrics*, 29, 17-22.

Brownlee, K.A. (1965), *Statistical Theory and Methodology in Science and Engineering* (2nd ed.), New York: John Wiley.

Casella, G., and George, E. (1992), "Explaining the Gibbs Sampler", *The American Statistician*, 46, 167-174.

Cheng, R.C.H., and Amin, N.A.K. (1979), "Maximum Product of Spacings Estimation With Application to the Lognormal Distribution," Mathematics Report 79-1, University of Wales, Cardiff, Dept. of Mathematics.

Cheng, R.C.H., and Iles, T.C. (1987), "Corrected Maximum Likelihood in Non-regular Problems," *Journal of the Royal Statistical Society*, Ser. B, 49, 95-101.

Cheng, R.C.H., and Traylor, L. (1995), "Non-regular Maximum Likelihood Problems" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 57, 3-44.

Daniel, C., and Wood, F.S. (1971), *Fitting Equations to Data*, New York: John Wiley.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

Fang, K.T., Kotz, S., and Ng, K.W. (1990), *Symmetric Multivariate and Related Distributions*, London: Chapman and Hall.

Fernández, C., and Steel, M.F.J. (1996a), "On the Dangers of Modelling Through Continuous Distributions: A Bayesian Perspective," mimeo, Tilburg University, CentER for Economic Research.

Fernández, C., and Steel, M.F.J. (1996b), "On Bayesian Modelling of Fat Tails and Skewness", Discussion Paper 9658, Tilburg University, CentER for Economic Research.

Fisher, F.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 8, 179-188.

Gantmacher, F.R. (1959), *The Theory of Matrices* (Vol.1), New York: Chelsea.

Gelfand, A., and Smith, A.F.M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities", *Journal of the American Statistal Association*, 85, 398-409.

Geweke, J. (1993), "Bayesian Treatment of the Independent Student-t Linear Model," *Journal of Applied Econometrics*, 8, S19-S40.

Geweke, J. (1994), "Priors for Macroeconomic Time Series and Their Applications," *Econometric Theory*, 10, 609-632.

Giesbrecht, F., and Kempthorne, O. (1976), "Maximum Likelihood Estimation in the Three-parameter Lognormal Distribution", *Journal of the Royal Statistical Society*, Ser. B, 38, 257-264.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations*, New York: John Wiley.

Heitjan, D.F. (1989), "Inference From Grouped Continuous Data: A Review" (with discussion), *Statistical Science*, 4, 164-183.

Kong, A., Liu, J.S., and Wong, W.H. (1994), "Sequential Imputations and Bayesian Missing Data Problems," *Journal of the American Statistical Association*, 89, 278-288.

Lange, K.L., Little, R.J.A., and Taylor, J.M.G. (1989), "Robust Statistical Modeling Using the t-Distribution. *Journal of the American Statistical Association*, 84, 881-896.

Lange, K.L., and Sinsheimer, J.S. (1993), "Normal/Independent Distributions and Their Applications in Robust Regression," *Journal of Computational and Graphical Statistics*, 2, 175-198.

Lehmann, E.L. (1983), *Theory of Point Estimation*, New York: John Wiley.

Liu, C.H. (1995), "Missing Data Imputation Using the Multivariate t Distribution," *Journal of Multivariate Analysis*, 53, 139-158.

Liu, C.H. (1996), "Bayesian Robust Multivariate Linear Regression With Incomplete Data," *Journal of the American Statistical Association*, 91, 1219-1227.

Liu, C.H., and Rubin, D.B. (1994), "The ECME Algorithm: A Simple Extension of EM and ECM With Faster Monotone Convergence," *Biometrika*, 81, 633-648.

Liu, C.H., and Rubin, D.B. (1995), "ML Estimation of the Multivariate t Distribution With Unknown Degrees of Freedom," *Statistica Sinica*, 5, 19-39.

Maronna, R. (1976), "Robust M-estimators of Multivariate Location and Scatter," *Annals of Statistics*, 4, 51-67.

Murray, G.D. (1977), Comment on "Maximum Likelihood From Incomplete Data Via the EM Algorithm," by A.P. Dempster, N.M.Laird, and D.B. Rubin, *Journal of the Royal Statistical Society*, Ser. B, 39, 27-28.

Pettitt, A.N. (1985), "Re-Weighted Least Squares Estimation With Censored and Grouped Data: An Application of the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 47, 253-260.

Relles, D.A., and Rogers, W.H. (1977), "Statisticians Are Fairly Robust Estimators of Location," *Journal of the American Statistical Association*, 72, 107-111.

Resek, R.W. (1976), "Estimation of the Parameters of a General Student's t Distribution," *Communications in Statistics, Part A-Theory and Methods*, 5, 635-645.

Rousseeuw, P.J., and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633-639.

Rubin, D.B. (1987), "A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Observations Are Modest: The SIR Algorithm, Comment on Tanner and Wong (1987)," *Journal of the American Statistical Association*, 82, 543-546.

Smith, R.L. (1989), "A Survey of Nonregular Problems," in *Proceedings of the International Statistical Institute Conference, 47th Session*, pp.353-372.

Tanner, M.A., and Wong, W.H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528-550.

Tiede, J.J., and Pagano, M. (1979), "The Application of Robust Calibration to Radioimmunoassay," *Biometrics*, 35, 567-574.

West, M. (1984), "Outlier Models and Prior Distributions in Bayesian Linear Regression," *Journal of the Royal Statistical Society*, Ser. B, 46, 431-439.

Zellner, A. (1962), "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, 58, 977-992.

Figure 1(a): Posterior Inference Through Set Observations

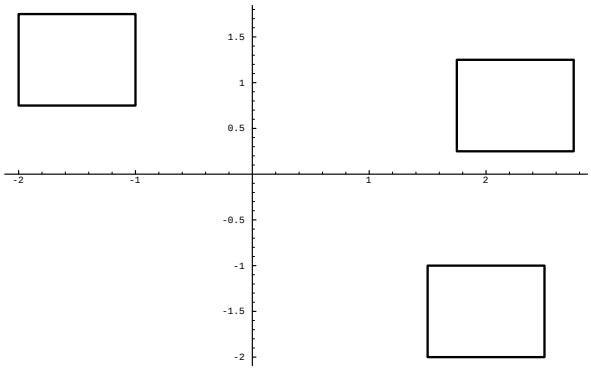
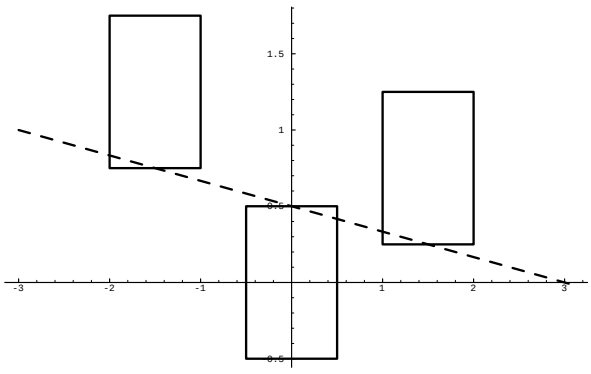


Figure 1(b): No Posterior Inference Through Set Observations



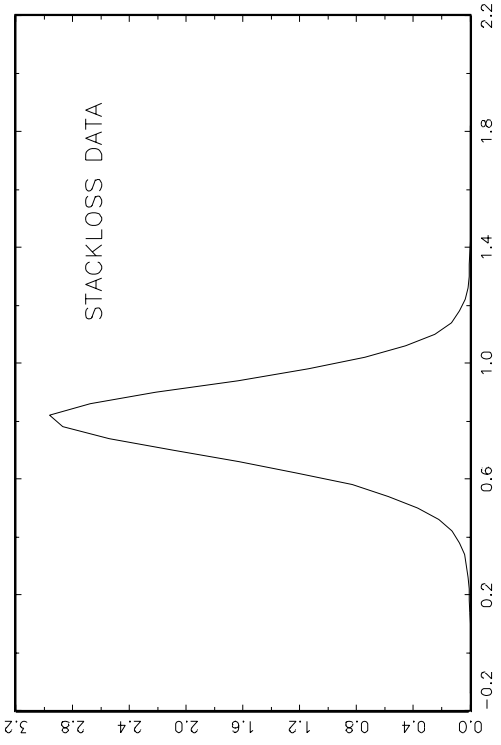


Figure 2: Posterior Density of beta2

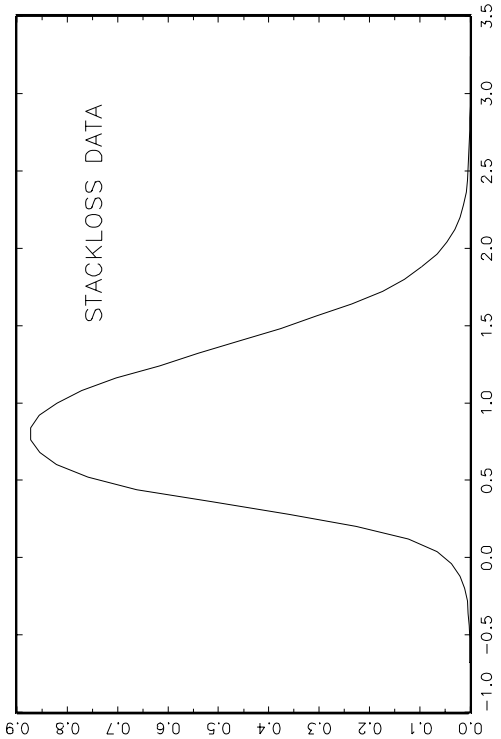


Figure 3: Posterior Density of beta3

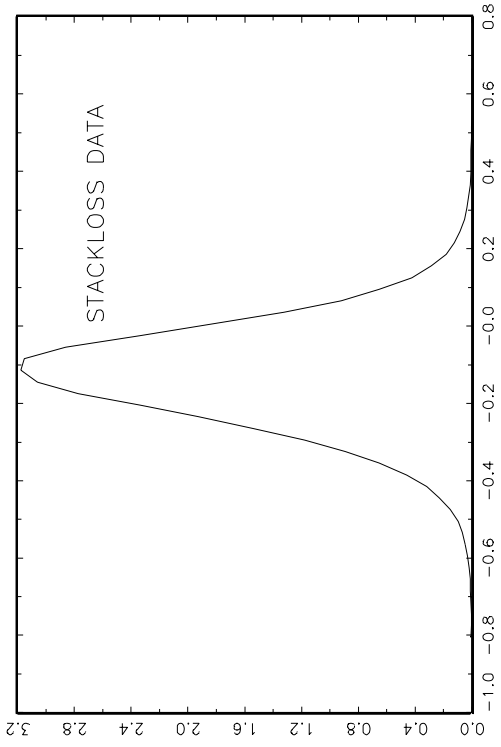


Figure 4: Posterior Density of beta4

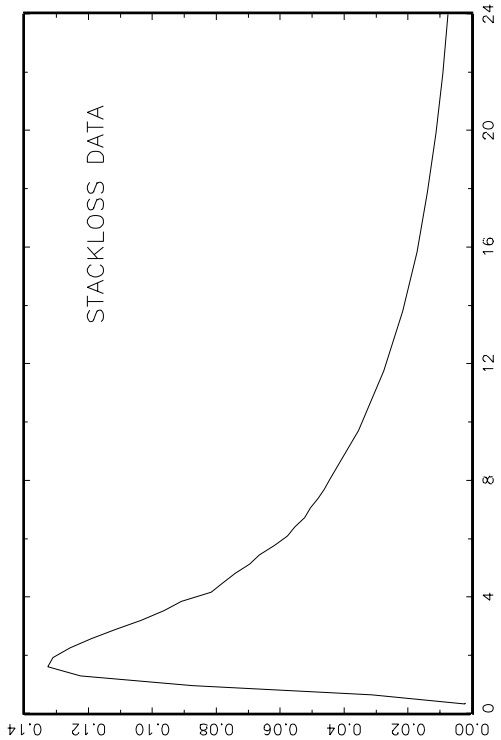


Figure 5: Posterior Density of nu

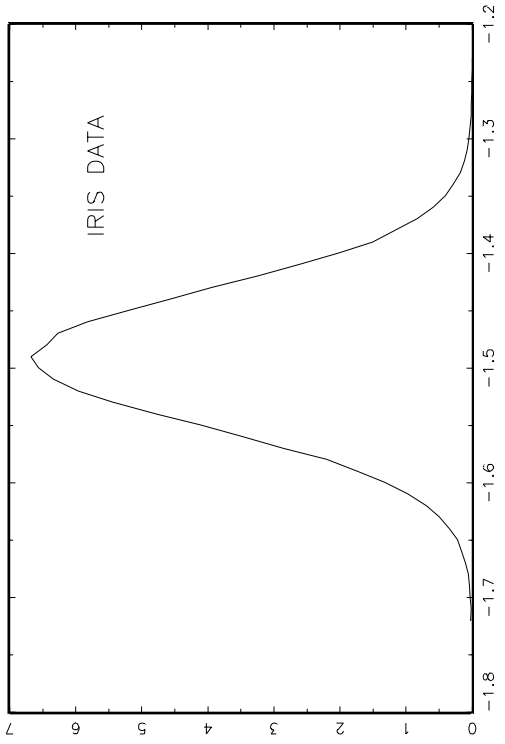


Figure 7: Posterior Density of beta2

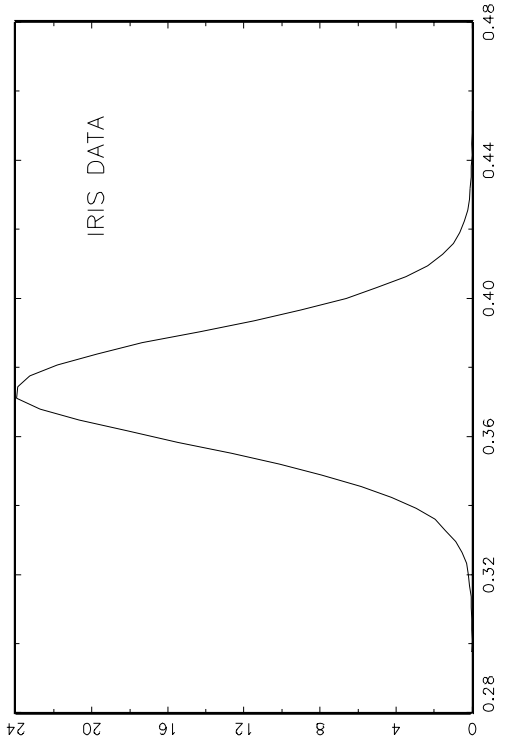


Figure 6: Posterior Density of beta1

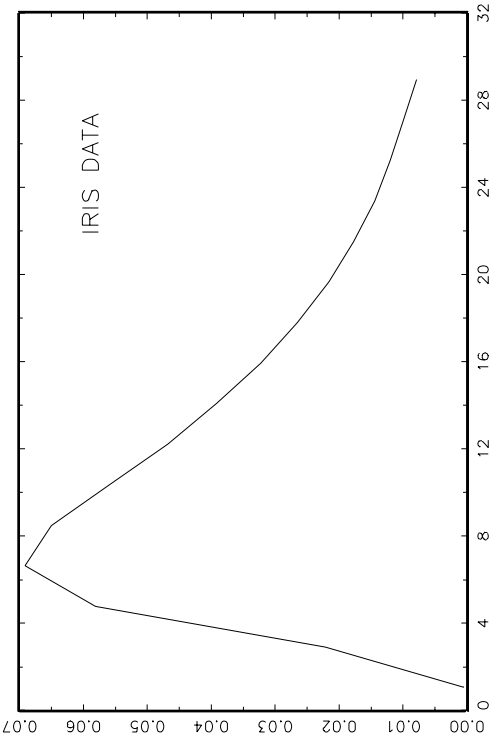


Figure 9: Posterior Density of nu

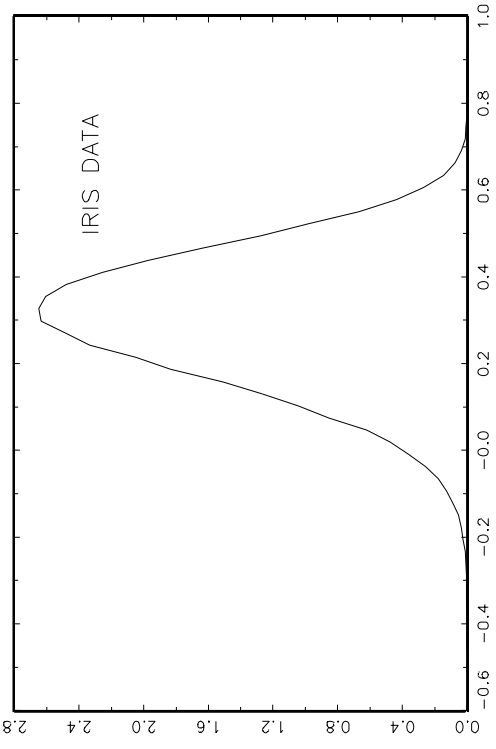


Figure 8: Posterior Density of correlation

Figure 10: Set Observations With Missing Components Do Not Help

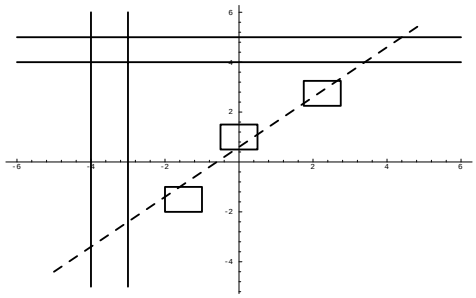
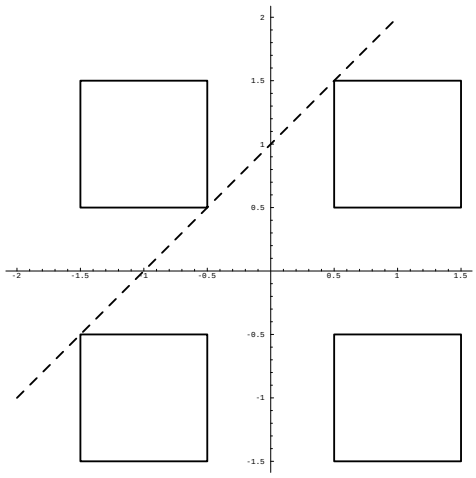


Figure 11: Compact Set Observations in Murray Data



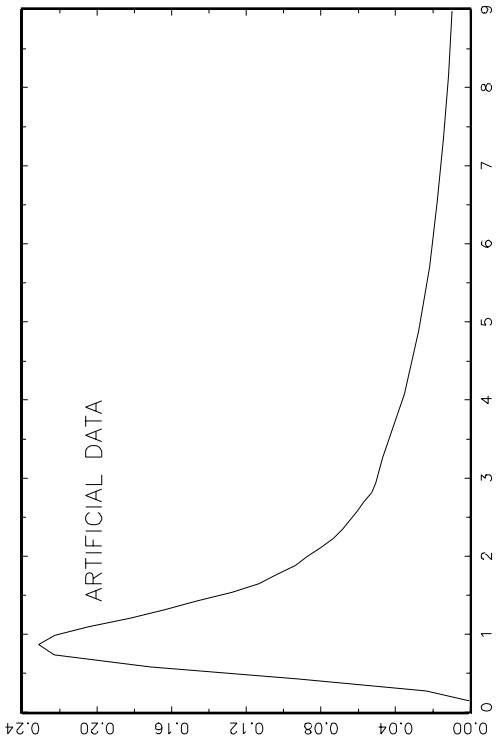


Figure 13: Posterior Density of nu

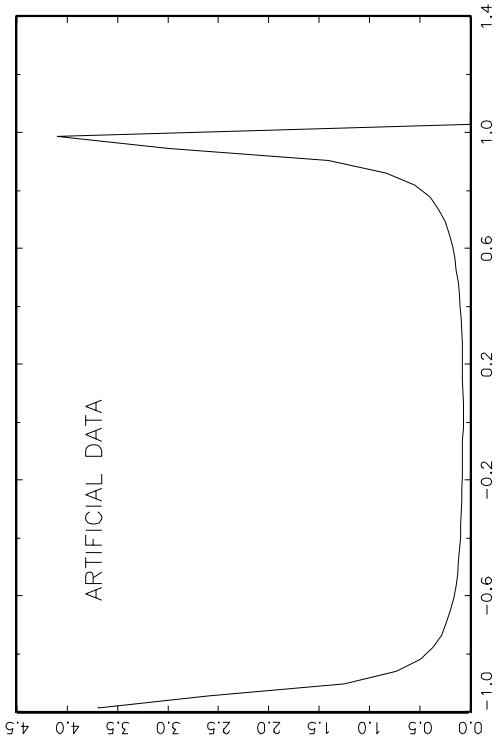


Figure 12: Posterior Density of correlation

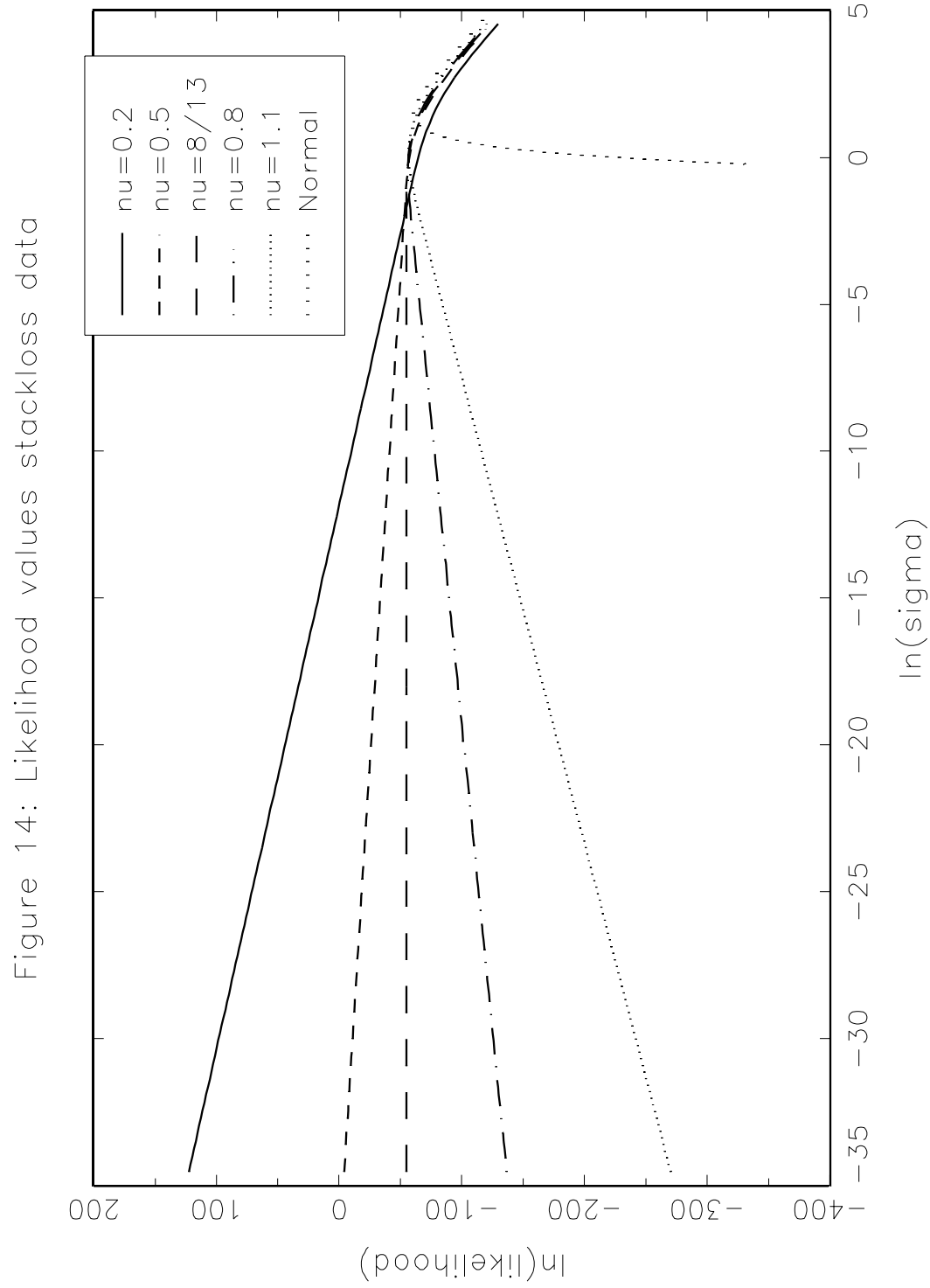


Figure 14: Likelihood values stackloss data